学校编码：10384　　分类号_____密级_____
学号：24320101152249　　UDC _____

厦门大学

硕 士 学 位 论 文

医疗数据的离群点检测方法研究

Research on Medical Data Outlier Detection Approaches

黄 艳 艳

指 导 教 师：廖 明 宏　教 授

专 业 名 称：计算机软件与理论

论文提交日期：2013 年 4 月

论文答辩日期：2013 年 6 月

学位授予日期：2013 年 6 月

指 导 教 师：_____

答辩委员会主席：_____

2013 年 6 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

离群点检测是数据挖掘领域一个重要的研究方向，用于揭示隐藏在数据中的重要信息，尤其在医疗诊断，入侵检测网络，信用卡欺诈，传感器敏感事件检测，地球科学等领域被广泛应用。而由于人眼只擅长处理二维或者三维的可视数值数据，所以利用人眼发现高维数据集中的离群点往往是比较困难的。因此我们对离群点检测技术深入研究是必要的。

针对医疗诊断数据的离群点检测方法进行了深入分析与仿真实验研究，取得了具有理论意义和应用价值的结果。

首先，针对给定的医疗数据集中离群点显著地偏离数据集中的其余对象，本文提出一种基于平均距离和平均密度的离群点检测的改进方法。该方法对数据集进行全局离群点检测，并利用平均距离求得每个数据对象的平均密度，随后计算平均邻域邻居数进行数据划分，最后用选择算法对数据对象进行筛选。

其次，针对数据集中离群点局部远离数据集的大多数，本文提出了一种基于图论的离群点检测的改进方法。该方法使用平均距离作为权重来度量每个数据对象的累积入度值，并通过特定的阈值 T 来划分疑似离群点数据集，最后用选择算法对数据对象进行筛选。

第三，针对数据集离散分布的特点，本文提出了一种基于 K-S 双样本检验的离群点检测的改进方法。该方法基于采用两个新的累积分布函数来检验，分别是来自数据集同一数据对象的 K 个最邻近距离的累积分布函数以及这 K 个点的两两距离的累积分布函数，分析它们的相似程度来判断它们是否属于同一分布。

仿真实验表明，三种离群点算法的改进都能针对特定的医疗数据集进行准确度为80%以上的离群点检测，并维护一个较低的误检率，其综合性能适用于数据挖掘的应用。

关键字：离群点检测；医疗数据；数据挖掘

Abstract

Outlier detection is an important research work in the field of data mining used to reveal important information hidden in the data. It has been widely used especially in medical diagnosis, intrusion detection network, credit card fraud, interesting sensor events, earth science and other fields. Because the human eye is only good at processing two-dimensional or three-dimensional visualization of numerical data, so it's more difficult to find outliers from high-dimensional data set by the use of the human eye. Therefore, the study of outlier detection technology research is necessary.

This paper discusses outlier detection method for medical diagnosis data. With in-depth analysis and simulation experiment research, the paper obtained both theoretical and application valuable result.

First of all, this paper proposes an outlier detection approach based on the average distance and the average density for a given medical data set which outliers deviate significantly from the rest of the objects of the data set. The approach of outlier detection is for global data sets, and takes advantage of the average distance to obtain the average density of each data object, and then calculates the average number of neighbors and divides the data objects into two different data subsets, and finally selects outliers from the candidate set.

Secondly, the paper proposes an outlier detection approach based on graph for a data set where local outliers are away from the majority of the data set. This approach takes advantage of the average distance to measure the cumulative in degree of each data object as a weight, and to divide the candidate data set through a specific threshold T, and finally with the selection algorithm for screening.

Thirdly, the paper proposes an outlier detection approach based on two-sample KS test for the discrete distribution characteristics of data sets. The approach is based on the use of two new cumulative distribution functions to verify the cumulative distribution function if the same data object. The two new cumulative distribution functions are the same data object but from the cumulative distribution function of the K nearest neighbors' distances and the cumulative distribution function of K pairwise

distances. This uses to determine whether they belong to the same distribution.

Simulation results show that the three outlier algorithm can be for some specific medical data sets accuracy of more than 80% of outlier detection, and maintain a low false alarm rate. Its overall performance is suitable for the application of data mining.

Keyword: Outlier Detection; Medical Data; Data Mining

目　录

CONTENTS

第 1 章 绪论

1.1 选题背景及研究目的和意义

离群点检测（outlier detection），也称为异常检测（anomaly detection），是数据挖掘领域一个重要的研究方向，用于发现不具备数据一般特性的数据对象。

数据挖掘的大多数工作集中在发现数据集合的“大模式”，如分类[1]、聚类[2]和关联规则等挖掘方法。那些不符合大多数数据对象所构成的规律的数据对象，则被称为离群点，或者噪声，或者意外排除在数据挖掘的分析处理范围之内，它们与数据的其它部分不同或者不一致[3]。

在很多应用领域中，数据的创建经过一个或多个阶段，往往能收集活动信息的观察值或反映系统的活动情况。当数据以一种不寻常的方式生成时，便是创建异常数据的过程。因此，离群点可能揭示隐藏在数据中的重要信息，例如商业欺诈中，小概率事件往往比经常发生的事件更有挖掘价值。离群点检测正是挖掘数据集合中与一般数据模型不相符合的那些数据。目前，离群点检测已广泛应用在以下几个领域：

（1）医疗诊断[4-6]。在许多医疗应用中，数据是从诸如 MRI 扫描（磁共振成像）、PET 扫描（正电子发射型计算机断层显像）或者心电图的时间序列等各种各样的设备而得到。这些数据中不正常的模式通常会反映疾病的状况。

（2）入侵检测网络[7, 8]。在很多情况下，基于主机或网络计算机系统，数据收集来自不同的类型，包括操作系统调用、网络流量或系统中的其他活动。由于恶意的活动，数据可能会显示不正常的行为。这种活动检测就被称为入侵检测。

（3）信用卡欺诈[9, 10]。目前信用卡诈骗已是相当普遍，因为如信用卡卡号这种敏感信息比较容易获取到，而这通常会导致未经授权就可以使用信用卡。在多数情况下，未经授权的使用可能会显示不正常的数据模式，如在某个隐蔽地域的大单交易。这样的模式可以用来检测在信用卡交易数据的异常值。

（4） 传感器敏感事件检测[11]。传感器通常是在实际应用中用来追踪各种环境和位置参数，在这种基本模式下的突然变化有可能代表感兴趣的事件。事件检测是传感器网络领域中一个主要的推动应用。

（5）地球科学[12]。大量的时空数据如天气模式、气候变化、土地覆盖格局等通过各种机制如卫星或遥感收集，在这些数据中异常往往会揭示导致这些异常出现的隐藏人类或者环境发展趋势信息。

正如前面所讲，异常检测在医疗、金融、社交、网络以及地球科学等领域应用，而且一般来说，现实中的数据都具有动态性、多样性和高维性。使用数据可视化方法来进行离群点检测如何？既然人眼在发现数据的不一致上非常迅速有效，这看起来可能是一个明显的选择[13]。但是，这不适用于高维大量的数据，离群点的值看上去实际可能是跟正常点完全一样有效的值，而且数据可视化方法对于检测有很多分类属性的数据，或高维数据中的离群点效果很差，这是因为人眼只擅长处理两到三维的可视数值数据,发现数据集中的离群点通常是比较困难的，因此我们对离群点检测技术深入研究是必要的。

1.2 国内外的研究现状及分析

近年来电话公司、信用卡公司、保险公司以及股票交易商对于诈欺行为的侦测（Fraud Detection）都很有兴趣，这些行业每年因为诈欺行为而造成的损失都非常可观，数据挖掘可以从一些信用不良的客户数据中找出相似特征并预测可能的诈欺交易，达到减少损失的目的。财务金融业可以利用数据挖掘来分析市场动向，并预测个别公司的营运以及股价走向。数据挖掘的另一个独特的用法是在医疗业，用来预测手术、用药、诊断、或是流程控制的效率。

在以前的很多研究中，往往将离群点的检测作为聚类算法的副产品，这些算法更多地认为离群点是嵌入到类中的背景噪声数据[14]。离群点检测是数据挖掘中一个重要的研究方面[13, 15, 16]，与关联规则挖掘、分类、类描述等数据挖掘不同，离群点发现用来检测数据集中小部分对象，即数据集中显著不同于其它数据的可疑对象。

研究人员根据对离群点存在形式的不同假设，提出了多种离群点检测算法，大致可划分为 4 大类别，其分别是：基于统计的方法、基于距离的方法、基于密度的方法、基于进化论的方法。

除了这些传统算法外，近几年，对离群点检测算法的研究取得了很多进展。例如针对高维数据的离群点检测，常用到的措施是将高维数据映射到低维子空间

来发现离群点；将基于距离和基于密度的算法并行化，能显著提高算法性能。近来，研究人员针对传统算法的种种缺陷开发了许多算法，能更有效，更高效地检测出离群点。

在国外离群点检测获得了广泛的研究和应用，E.M.Knorr 和 R.T.Ng[17]将离群点检测用于分析 NHL(National Hockey League)的运动员统计数据，用来发现表现特殊的运动员:K. Yalnanishi 和 J. Takeuchi[18]演示了如何将离群点检测应用于股票数据的变动检测;J. Laurikkala，M. Juhola[4]等研究了离群点在医学数据中的应用；L.GroSS 则将离群点分析应用于数据质量评价；空间离群点检测被大量研究；在数据仓库领域，离群点检测被用来发现不一致的数据，提高数据质量。

将数据点的 k 最近邻距离作为离群程度指标能够有效发现数据集中的离群点，但是基本算法需要 $O(n^2)$次数据点间的距离计算，不适用于大数据集，为此邵纪东等人[19]提出了利用度量空间三角不等式的快速挖掘算法—提前修剪(ADVP)。

针对 LOF 算法存在的不同问题，Ada Wai-chee Fu 和 Anny Lai-mei Chiu[20]对算法进行三种改进。分别给出两个新的离群程度定义，并提出了 GirdLOF 算法用于 LOF 算法之前的修剪.与 LOF 算法相比，这三种改进算法都在不同方面有所提高。

为了使离群点检测更为自动化，减少用户对参数选择的困难性，施化吉等人[21]提出了平均密度的定义，并给出了基于平均密度的离群点检测方法。该方法不仅能够有效地检测出离群点，同时也简化了离群点检测时对用户输入参数的要求。

Hans-Peter Kriegel 等人提出了 ABOD 算法[22]，即基于角度的离群点检测算法。算法利用角度作为度量，在聚类中心的点与其他点形成的角度和会变小，离群点与其他点形成的角度和最小。

基于单元的离群点检测算法适用于低维情况，当维数增加以及划分粒度较细时，由于生成单元数多，算法不能有效地工作，针对这个问题，孙焕良等人[23]设计了一种新的索引结构 CD-Tree，并给出了基于划分的数据偏斜度概念，在此基础上提出了一种基于划分的离群点检测算法，该算法相比基于单元的算法，在效率及有效处理的维数方面均有显著提高。

HiOut 算法[24]由 Angiulli 和 Pizzuti 提出。Aggarwal 和 Yu 开发了基于稀疏性系数的子空间离群点检测方法[25]。

1.3 本文主要研究内容及结构

本论文共分为六章，结构如下。

第一章为绪论，介绍了离群点检测的研究现状及分析，以及本文的主要工作。

第二章介绍了离群点检测问题的定义，总结了离群点检测算法的分类方法，针对其中关于正常对象和离群点的假定，对各方法分组，详细介绍了统计学的方法，基于邻近性的检测方法，基于聚类的方法，以及高维数据中的离群点检测方法。

本文的主要研究成果集中在第三章和第五章。针对医疗数据的离群点检测，本文提出了三种改进算法。

第三章研究了一种混合的基于距离的离群点检测方法，提出了一种基于平均距离和平均密度的离群点改进检测方法 DDOD(Distance- Density Outlier Detection)。该算法利用平均领域邻居数目来划分疑似离群点和非离群点的数据集，在筛选阶段，通过选择算法筛选 K 邻近距离较大的点从而检测出离群点。仿真表明，改进的算法 DDOD 的检测率大部分达到 90%以上，甚至达到 100%。

第四章研究了一种基于图论的离群点检测方法，提出了一种基于入度值的离群点改进检测方法 INOD(In Degree Outlier Detection)。该算法使用平均距离作为权重来度量每个数据对象的累积入度值，并通过特定的阈值 T 来划分疑似离群点数据集，最后用选择算法对数据对象进行筛选。仿真表明，改进的算法 INOD 的检测率针对特定数据集可以达到 100%，并且维护一个较低的误检率。

第五章研究了一种基于统计学的离群点检测方法，提出了一种基于 K-S 双样本检验的离群点检测方法 KSOD(K-S test Outlier Detection)。该方法采用两个新的累积分布函数来检验，分别是来自数据集，同一数据对象的 K 个最邻近距离的累积分布函数以及这 K 个点的两两距离的累积分布函数，分析它们的相似程度来判断它们是否属于同一分布，从而检测出离群点。仿真表明，改进的方法 KSOD 离群点检测方法，使用曼哈顿度量距离，针对特定数据集得到的离群点检测率可

以达到 100%。

第六章是本论文的结束语，总结了前面各章的研究内容，并讨论了未来的研究。本文得出的三种改进的离群点算法的改进都能针对特定的医疗数据集进行准确度为 80%以上的离群点检测，并维护一个较低的误检率，其综合性能适用于数据挖掘的应用。

第 2 章 离群点检测技术分析

2.1 离群点的定义

离群点是那些不符合大多数数据对象所构成的规律的数据对象[3, 26]。在图 2-1 中，大部分对象都粗略地服从高斯分布。然而，区域 *R* 中的对象显著不同。它不太可能与数据集中的其他对象服从相同的分布。因此，在该数据集中，*R* 中的对象是离群点。

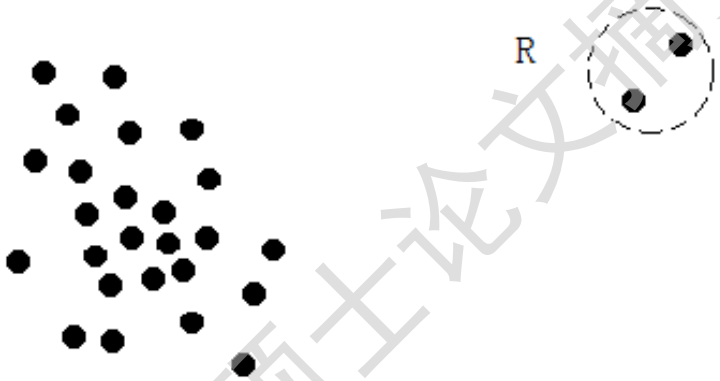

图 2-1 区域 R 中的对象是离群点

离群点不同于噪声数据。噪声是被观测变量的随机误差或方差。一般而言，噪声在数据分析包括离群点分析中不是令人感兴趣的。与许多其他数据分析和数据挖掘任务一样，应该在离群点检测前就删除噪声。

离群点的成因主要有三种形式。

第一种是由错误产生的离群点，如不完善的数据采集设备或手工输入时丢失弄错数据、测量单位混乱，数据传输错误和机器故障造成的数据异常等。

第二种是由数据变异产生的离群点，这是数据分布真实性的反映。

第三种是数据来源于不同的类。定义给定的数据集是来源于不同的类，其中较少分类的数据被认为是离群点数据。如公司经理的工资与一般员工相比可能就是一个异常数据。